Comparative corpus linguistics: new perspectives and applications

Convenors:
Natalia Levshina (Leipzig University, Leipzig)
Annemarie Verkerk (Max Planck Institute for the Science of Human History, Jena)
Steven Moran (University of Zurich, Zurich)

Although the main bulk of existing corpus-based research is probably formed by language-specific descriptive studies, corpora have long been used successfully for large-scale language comparison and for testing linguistic generalizations, e.g. Zipf (1935) and Greenberg (1960). Nowadays, linguists can enjoy the abundance of large comparable and parallel corpora and other multilingual resources, such as the Universal Dependencies Corpora (Nivre et al. 2017), the parallel Bible translations (Mayer & Cysouw 2014), OPUS corpus (Tiedemann 2012), Multi-CAST (Haig & Schnell 2016) and Google Books Ngrams. The availability of such resources provides functional linguists, typologists, historical linguists and psycholinguists with new exciting opportunities to answer big theoretical questions, exemplified by successful applications of comparative corpus-based approaches such as the following:

- formulation, refinement and explanation of linguistic generalizations, e.g. Zipf's Law of Abbreviation (Piatandosi et al. 2011; Bentz & Ferrer-i-Cancho 2016), the principle of dependency length minimization (Futrell et al. 2015) and the principle of economy in morphosyntactic alternations (Haspelmath et al. 2014);
- computation of corpus-based measures that represent typological parameters, such as analyticity, syntheticity, complexity or referential density (e.g. Juola 1998; Bickel 2003; Stoll & Bickel 2009; Szmrecsanyi 2009; Ehret & Szmrecsanyi 2016);
- using massively parallel and comparable corpora for unsupervised pattern detection, e.g. finding the universal conceptual dimensions of motion verbs (Wälchli & Cysouw 2012) and automatic extraction of typological features (Virk et al. 2017);
- development of new statistical methods, and probabilistic and connectionist approaches to the study of language acquisition (e.g. Chater & Manning 2006, Behrens 2008), in particular from a cross-linguistic perspective (MacWhinney & Snow 1985; Moran et al 2016);
- quantitative diachronic typology, e.g. development of manner and path verbs in Indo-European (Verkerk 2015);
- detection of areal patterns in genealogically related languages (e.g. van der Auwera et al. 2005; von Waldenfels 2015);
- usage-based explanations of the evolution of linguistic types, e.g. studies related to the Preferred Argument Structure hypothesis (Du Bois 1987; Haig & Schnell 2016);
- cross-linguistic comparison of probabilistic constraints on multifactorial language variation, e.g. the use of analytic and lexical causatives (Levshina 2016).

The aim of this workshop is to bring together typologists, functional linguists, psycholinguists and other specialists who use cross-linguistic corpora for testing their hypotheses, and corpus linguists who build and use such corpora to address research questions in linguistic diversity. We want to discuss the recent developments, perspectives and challenges of corpus-based language comparison. We seek contributions that sample a sizable amount of the world's languages or language varieties, whether at the global level, or

within particular families or areas. A list of potential contributions includes, but is not limited to, the following:

- case studies showing how one can use the information derived from corpora for the purposes of typological classification;
- corpus investigations of linguistic generalizations and explaining these findings in terms of processing-related, communicative and learning constraints or biases;
- corpus-based language comparison from a genealogical and/or areal perspective;
- corpus-based studies in diachronic typology and historical linguistics;
- studies addressing the problem of comparative concepts (Haspelmath 2010) and its consequences for comparative corpus linguistics, in particular, for the development of cross-linguistic annotation schemas;
- presentation of newly developed cross-linguistic corpora, preferably with a case study revealing their possibilities;
- discussion of statistical methods and visualization tools for analysing cross-linguistic corpus data.

References

Behrens, H. (ed.). (2008). *Corpora in language acquisition research: History, methods, perspectives* (Vol. 6). Amsterdam: John Benjamins.

Bentz, Ch., & Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In Bentz, Christian, Gerhard Jäger and Igor Yanovich (eds.), *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. University of Tubingen, online publication system: https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558.

Bickel, B. (2003). Referential density in discourse and syntactic typology. *Language* 79, 708–736.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10(7), 335-344.

Du Bois, John W. (1987). The discourse basis of ergativity. *Language*, 64, 805–55.

Ehret, K. & Szmrecsanyi, B. (2016). An information-theoretic approach to assess linguistic complexity. In R. Baechler & G. Seiler (eds.), *Complexity and Isolation*, 71-94. Berlin: de Gruyter.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33),10336–10341.

Haig, G. & Schnell, S. (eds.). (2016). Multi-CAST (Multilingual Corpus of Annotated Spoken Texts), https://lac.uni-koeln.de/multicast/.

Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, *86*(3), 663–687.

Haspelmath, M., Calude, A., Spagnol, M., Narrog, H., & Bamyaci, E. (2014). Coding causal–noncausal verb alternations: A form–frequency correspondence explanation. *Journal of Linguistics*, *50*, 587–625. http://doi.org/doi:10.1017/S0022226714000255

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language. *International Journal of American Linguistics*, 26(3), 178–94.

Juola, P. (1998). Measuring linguistic complexity: the morphological tier. *Journal of Quantitative Linguistics*, 5(3), 206–213.

Levshina, N. (2016). Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages. *Folia Linguistica*, 50(2), 507–542.

Mayer, T., & Cysouw, M. (2014). Creating a massively parallel Bible corpus. *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, 3158–3163.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*(2), 271-295.

Moran, S., Schikowski, R., Pajovic, D., Hysi, C., & Stoll, S. (2016). The ACQDIV Database: Min (d) ing the Ambient Language. In *Proceedings of* the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 23 May 2016 - 28 May 2016, 4423-4429.

Nivre, J. Agić, Ž., Ahrenberg, L. et al. (2017). Universal Dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University in Prague, http://hdl.handle.net/11234/1-1983.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(9), 3526–3529.

Stoll, S. & Bickel, B. (2009). How deep are differences in referential density? In: Guo, J., Lieven, E., Budwig, N. et al. (eds.), *Crosslinguistic Approaches to the Psychology of Language*, 543–555. London: Psychology Press.

Szmrecsanyi, B. (2009). Typological parameters of intralingual variability: grammatical analyticity versus syntheticity in varieties of English. *Language Variation and Change*, 21(3), 319–353.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*

van der Auwera, J., Schalley, E., & Nuyts, J. (2005). Epistemic possibility in a Slavonic parallel corpus: A pilot study. In B. Hansen & P. Karlik (eds.), *Modality in Slavonic Languages. New Perspectives*, 201–217. München: Sagner.

Verkerk, A. (2015). Where do all the motion verbs come from? The speed of development of manner verbs and path verbs in Indo-European. *Diachronica*, 32(1), 69-104.

Virk, Sh. M., Borin, L., Saxena, A. & Hammarström, H. (2017). Automatic Extraction of Typological Linguistic Features from Descriptive Grammars. In Kamil Ekštein & Václav Matoušek (eds.), *TSD 2017: Text, Speech, and Dialogue*, 111–119. Cham: Springer.

von Waldenfels, R. (2015). Inner-Slavic contact from a corpus driven perspective. In E. Kelih, S. M. Newerkla, & J. Fuchsbauer (eds.), *Lehnwörter im Slawischen: Empirische und crosslinguistische Perspektiven*, 237-263. Frankfurt: Peter Lang.

Zipf, G. K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Cambridge, MA: MIT Press.